Terabit-scale hybrid networking

Zhenzhen Yan, Tian Jin, M. Veeraraghavan
University of Virginia
mvee@virginia.edu

Chris Tracy
ESnet
ctracy@es.net

March 18, 2013
Project web site: http://www.ece.virginia.edu/mv/research/DOE09/index.html

Thanks to the US DOE ASCR for grants DE-SC0002350 and DE-SC0007341 (UVA) DOE for DE-AC02- 05CH11231 (ESnet)

Thanks to Chin Guok for work on QoS mechanisms for circuits
Thanks to Brian Tierney, Eric Pouyoul, Tareq Saif, Andy Lake for testbed
Thanks to Brent Draney, Jason Hick, NERSC, Yee-Ting Li, Wei Yang, SLAC, for GridFTP DTN login access

Outline

- Hybrid network traffic engineering system (HNTES) router config. (QoS)
  - alpha-flow identification and redirection
  - alpha flows: high-rate, large sized flows
  - threshold: 1 GB in 1 min
- NetFlow data analysis – new results
  - 4 routers
  - May-Nov. 2011 data analyzed

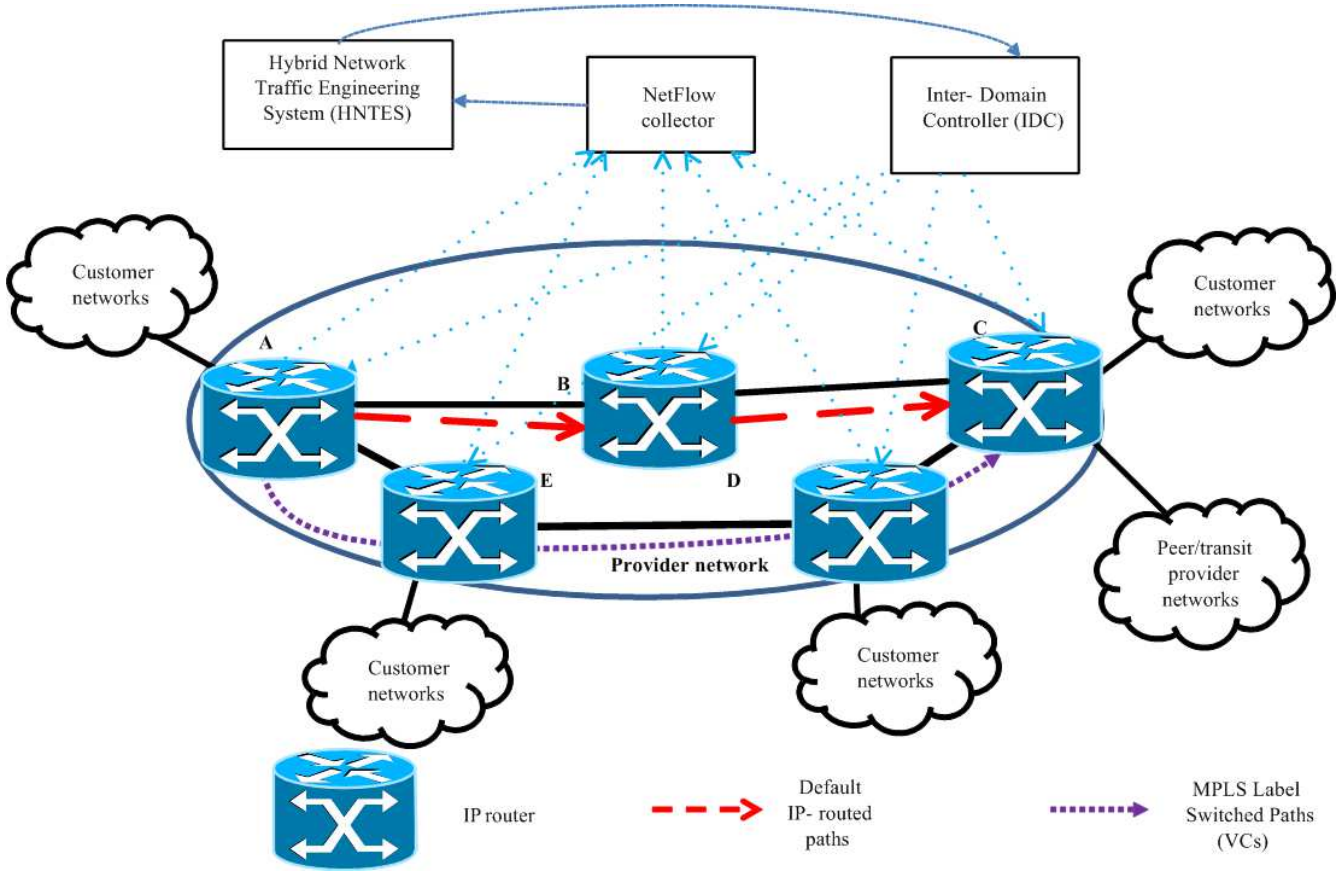
Two key findings

- Throughput on circuits could be lower than on IP-routed paths because of policing
  - IDC circuit provisioning includes policing
  - TCP not the culprit, but rather policing
  - Even with RoCE, throughput limited by circuit rate (which is typically less than link capacity)
- Online HNTES
  - Not required (4-router NetFlow analysis)
  - Impractical

Z. Yan, M. Veeraraghavan, C. Tracy, C. Guok, "On how to provision Quality of Service (QoS) for large dataset transfers," accepted in CTRQ 2013

Tian Jin, C. Tracy, M. Veeraraghavan, Z. Yan, "Traffic Engineering of High-Rate Large-Sized Flows," submitted to HPSR 2013

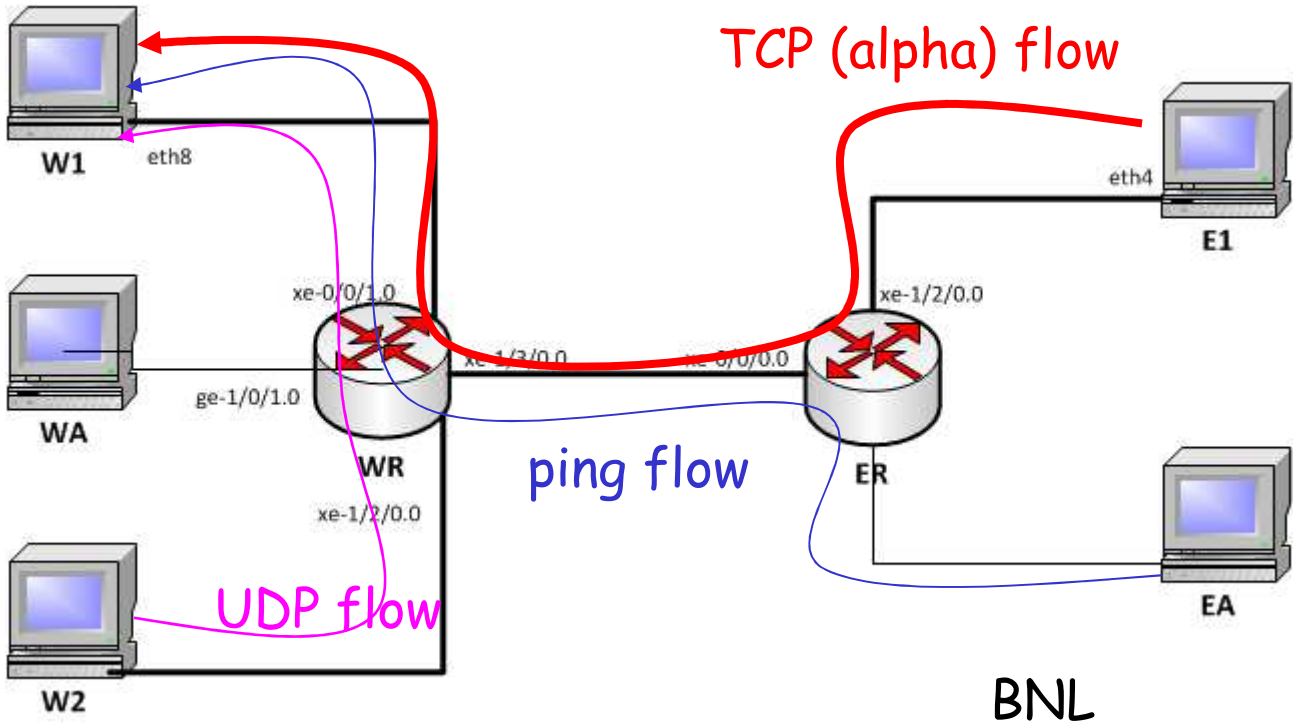
Hybrid network traffic engineering system (HNTES)
- Intradomain identification/redirection of alpha flows



- Three steps
  - offline (e.g., nightly) analysis of NetFlow data at ingress routers to identify address prefixes of alpha flows
  - requests L3 circuits between ingress-egress router pairs from IDC
  - IDC configures QoS mechanisms for circuit in routers & sets firewall filters to direct future packets with matching alpha prefix IDs to L3 circuits

Study router QoS configuration mechanisms

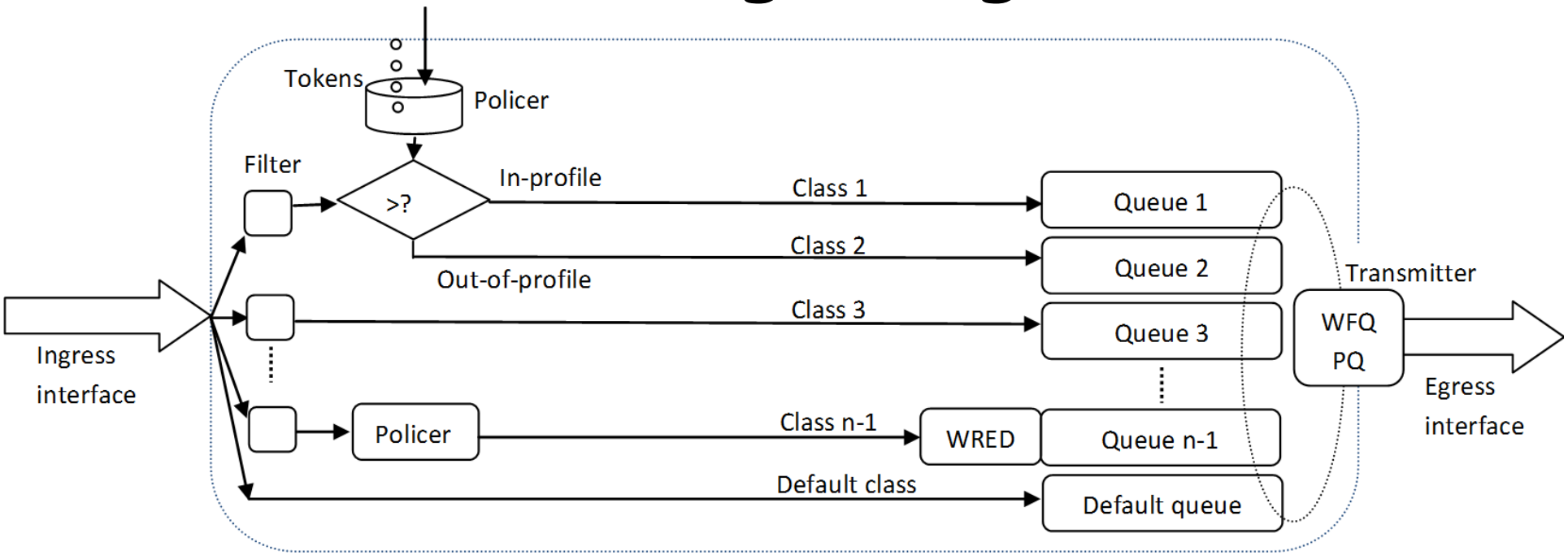
- Used DOE LIMAN testbed
- Hosts: high-performance diskpts



Router configurations

- Firewall filter:
  - alpha-flow based on /24 or /32 address prefixes (src and dst)
- Policing
  - classify out-of-profile packets as scavenger class and send to scavenger queue
  - Weighted Random Early Detection (WRED)
- Scheduling
  - Weighted fair queueing (WFQ)
  - Priority queueing (PQ)

Policing on ingress scheduling on egress



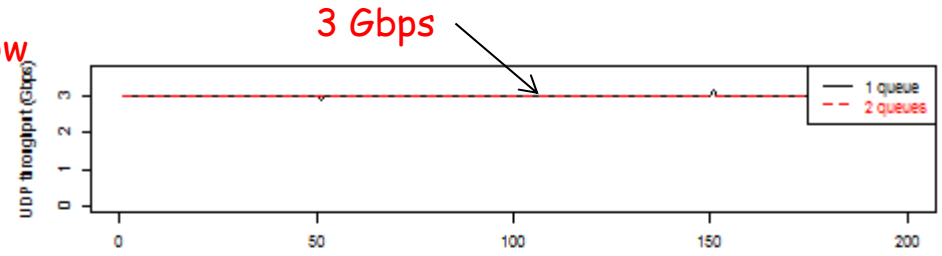
- Dual goals:
  - reduce impact of alpha flows on real-time delay-sensitive flows
  - allow alpha flows to enjoy high throughput

Compare 3 configurations

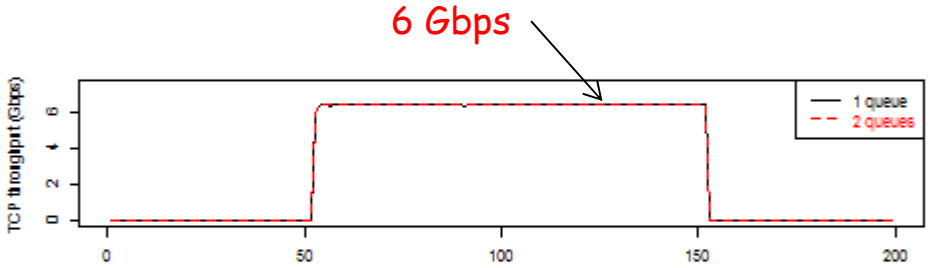
- 1-queue:
  - best-effort
  - all flows directed to same egress-side queue
- 2-queue: alpha and beta
  - scheduling-only (no policing)
    - WFQ + PQ: WFQ marks queues as in- or out-of-profile
    - transmitter: shared in work conserving mode (non-strict)
    - buffer: strict partitioning
- 3-queue: alpha, beta, scavenger service (SS)
  - policing: > 1 Gbps sent to SS queue
  - scheduling: WFQ + PQ

Impact of alpha flows on real-time flows

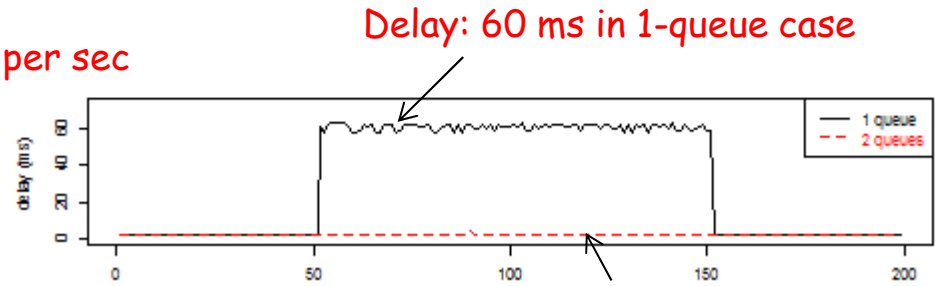
UDP flow



TCP flow

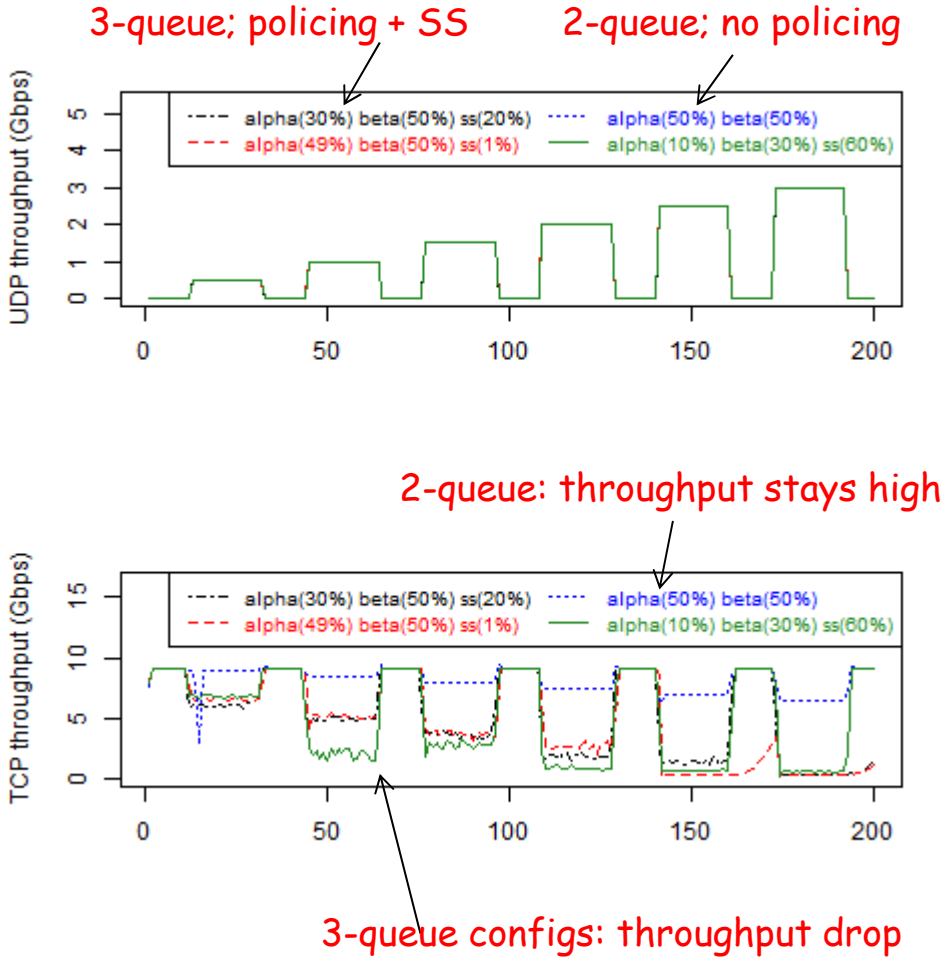


Pings: 1 per sec



- Impact on ping flow delay
  - significant in 1-queue configuration
  - negligible in 2-queue configuration
- Need separate virtual queue for alpha flow packets

Impact of policing: 3-queue case causes TCP throughput to drop

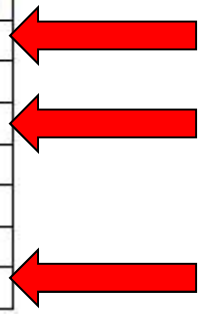


- When UDP rate is increased from 0.5 to 1 Gbps, significant drop in TCP throughput
- Why? out-of-sequence packets
- TCP fast-retransmit/fast recovery algorithm causes sending rate to drop by half
- Worst when alpha queue allocation is only 10%

Impact of WFQ settings

TABLE I: α-flow throughput under different background loads (UDP rate) and QoS configurations

UDP rate (Gbps)	α-flow throughput (Gbps)			
	Percentages for 2-queues (α, β) and 3-queues (α, β, SS) configurations			
	(50,50)	(49,50,1)	(30,50,20)	(10,30,60)
0	9.12	9.09	9.07	9.12
0.5	8.92	6.62	6.06	6.83
1	8.43	5.22	5	2.12
1.5	7.94	3.78	3.67	2.82
2	7.44	2.7	1.93	0.92
2.5	6.95	0.33	1.38	0.69
3	6.46	0.34	0.38	0.61



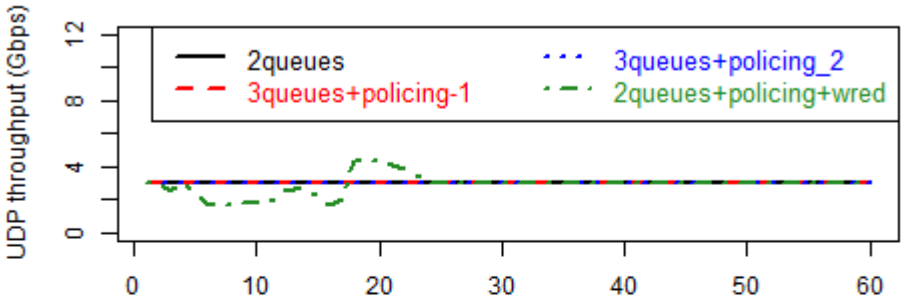
- Top row: For every 1 pkt in alpha queue, 9 in SS queue; WFQ will complete serving alpha queue and then serve SS queue packets; sequence preserved and throughput is high

- The higher the background traffic load, the lower the TCP-flow packet arrival rate to the policer, the larger the inter-arrival gaps, the higher the number of collected tokens in the bucket, and the larger the number of in-profile packets directed to the alpha queue.
- If the WFQ allocation to the alpha queue is insufficient to serve these in-profile bursts, packets from the alpha queue and SS queue will be intermingled resulting in out-of-sequence packets at the receiver and lowered throughput

WRED instead of SS queue

- In- and out-of-profile (OOP) packets are held in the same virtual queue
- Drop OOP packets according to a profile based on queue occupancy (e.g., increase linearly from 0 to 100 based on queue occupancy)
- WRED case does better than SS option, but still worse than no policing
- HTCP aggressive increase of cwnd

Compares no-policing, policing with SS, policing with WRED



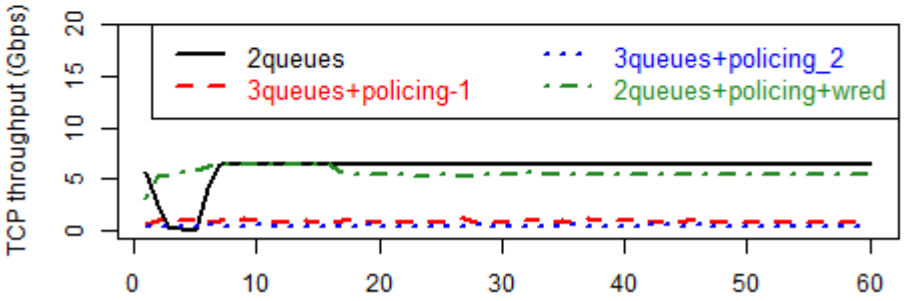


TABLE II: QoS configurations for experiment 3

Configuration	Policing	WFQ allocation 2-queues:(α,β) 3-queues:(α,β,SS)	WRED
2-queues	None	(60,40)	NA
3-queues + policing1	OOP to SS queue	(59,40,1)	NA
3-queues + policing2	OOP to SS queue	(20,40,40)	NA
2-queues + policing + WRED	WRED	(60,40)	Drop prob. = queue occ.

- Initial drop due to HTCP's aggressive increase of cwnd
- 125MB buffer
- 75MB for alpha Q (60-40)
- When TCP flow rate > 7 Gbps for short duration, Q fills up

Main point

- Dual goals
  - Separate out alpha flows to reduce their impact on delay/jitter of real-time flows
  - Let file transfers generating these alpha flows enjoy high throughput for reduced transfer times
- Both goals met with:
  - Rate-unspecified circuits
  - Separate Q for alpha flows but no policing

Outline

- Hybrid network traffic engineering system (HNTES)
  - alpha-flow identification and redirection
  - alpha flows: high-rate, large sized flows
  - threshold: 1 GB in 1 min
- NetFlow data analysis – new results
  - 4 routers
  - May-Nov. 2011 (214 days) data analyzed

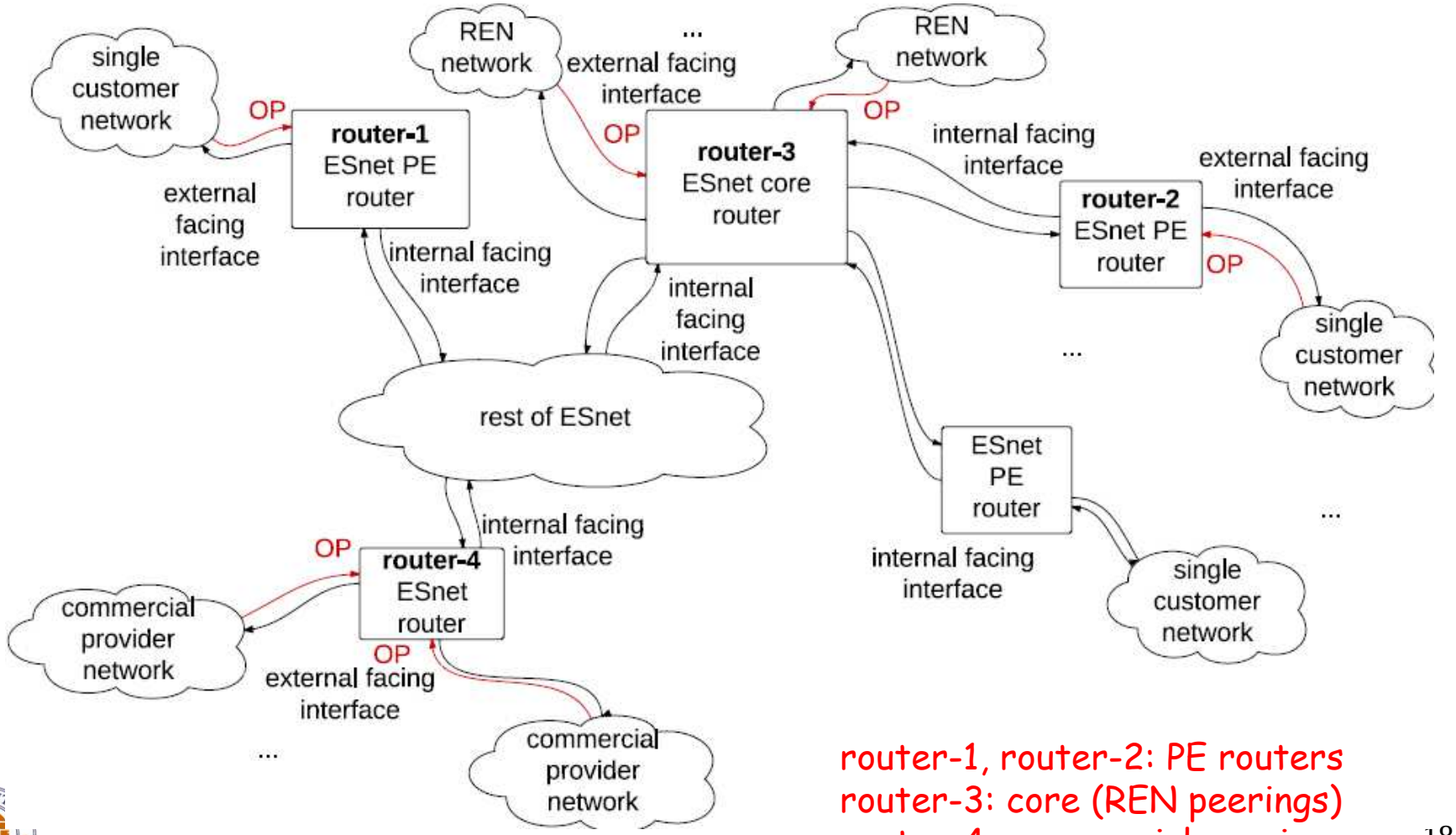
Purpose of analysis

- Is offline HNTES sufficient or is online HNTES required?
  - is effectiveness of offline HNTES low?
    - do new flows appear all the time?
  - is afflicted-flow pkt percent high?
    - need more specific firewall filters using port numbers not just addresses?
- Should we use /24 or /32 address prefixes in firewall filters?

Key findings

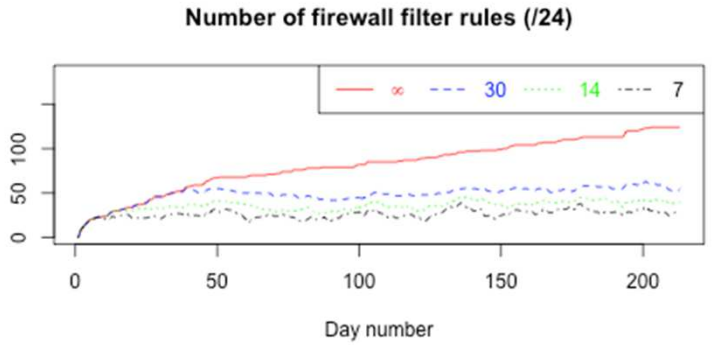
- Offline HNTES is sufficient
  - Effectiveness of offline scheme is high in PE routers – collect NetFlow samples in both directions of inter-domain links at PE routers
    - Same src-dest subnets repeatedly generate alpha flows
  - Afflicted-flow pkt percent is small
    - beta flows who share alpha flow prefix IDs
- Parameters
  - Use /24 prefixes instead of /32
  - Aging parameter: 30 days

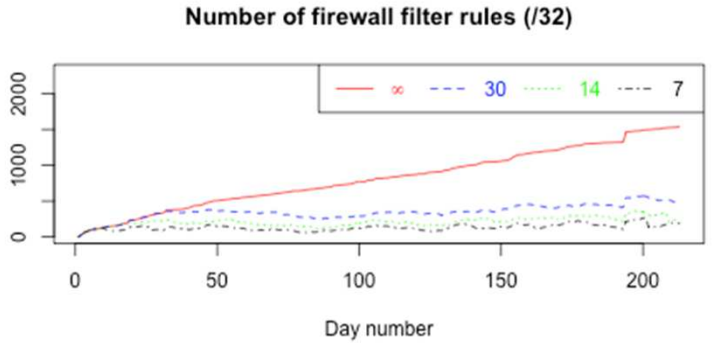
NetFlow observation points (OP)

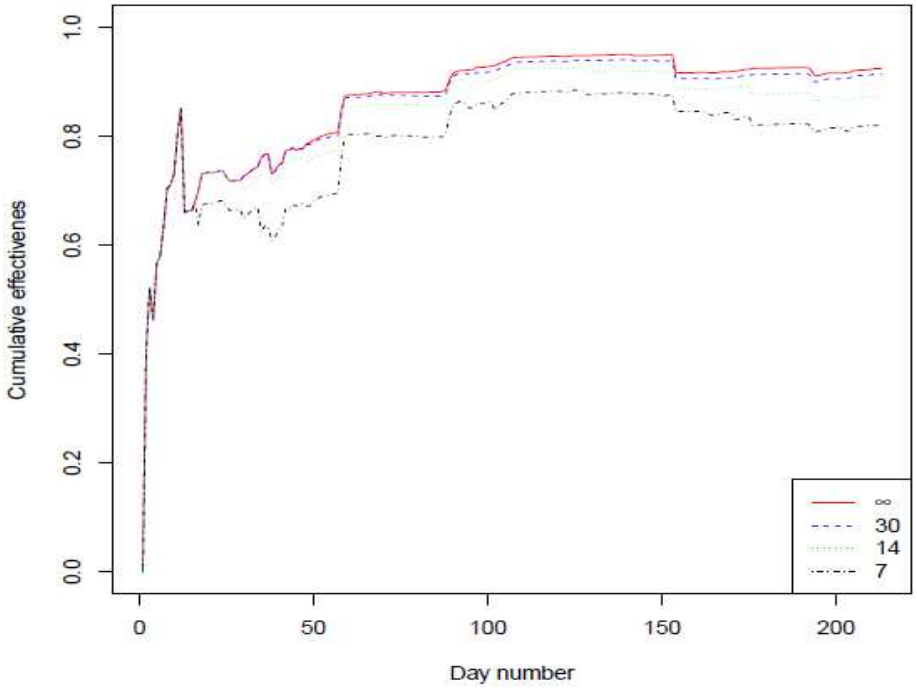


router-1, router-2: PE routers
router-3: core (REN peerings)
router-4: commercial peerings

Aging parameter: tradeoff effectiveness with size of firewall filter







- cumulative effectiveness: percent of alpha bytes that would have been redirected had HNTES been deployed
- 30 days is good compromise for aging parameter
- graphs for router 1; similar for other routers

Comparisons

TABLE I: Rows 1 – 3: across values from day 100 to day 214; Rows 4 – 8: across the whole 214-day period; The aging parameter A value is assumed to be 30 days (rows 7 and 8 are unaffected by the aging parameter)

Row	Statistics		router-1		router-2		router-3		router-4	
			/24	/32	/24	/32	/24	/32	/24	/32
1	Size of firewall filter	max	63	572	120	969	34	63	41	74
2		mean	53.41	406.77	91.63	384.32	24.63	48.82	29.36	8.4
3		cv	0.08	0.18	0.18	0.77	0.18	0.18	0.18	1.29
4	Cumulative effectiveness, C_{214}		91%	82%	92%	83%	83%	76%	67%	50%
5	# of days when $E_i = 1$		90	3	49	21	104	72	86	60
6	# of days when $E_i = 0$		1	5	2	4	12	23	25	51
7	# of days when no α flow appeared		1	1	0	0	21	21	35	35
8	total # of α prefix IDs		125	1548	281	1639	104	228	117	239

- Obs. 1: high effectiveness for router-1, router-2 because of repeated alpha flows between same src-dest pairs
- Obs. 2: lower effectiveness for router-3, router-4: fewer uploads to DOE sites than downloads
- Obs. 3: fewer alpha prefix IDs for router-3/4 but more days when daily effectiveness is 1 for /32 $\Rightarrow$ fewer servers involved in uploads than in downloads from DOE sites

Afflicted-flow packets

- Find B: set of non-alpha NetFlow reports that share alpha prefix IDs
- Create four subsets in sequence
  - C: set of reports that share 5-tuple IDs as alpha flows
  - D $\subseteq$ B-C: data-transfer reports (heuristic)
  - W $\subseteq$ B-C-D: well-known ports
  - L: leftover = B-C-D-W
- Afflicted flows: W+L

Afflicted-flow packets

TABLE IV: Percentage of afflicted-flow packets, AF_{214}

	router			
	1	2	3	4
/24	10.39%	23.84%	6.22%	25.37%
/32	11.22%	13.18%	3.43%	25.51%

- Small percentage of beta-flow packets that share alpha-flow prefix IDs will have adverse effects when redirected to alpha queue
- /24 vs /32
  - /32 has lower effectiveness: large % of beta-flow packets will be impacted when an alpha flow is not redirected
  - /24 has higher afflicted-flow %: small % of beta flows are adversely impacted

Key findings: HNTES

- Alpha-flow prefix identification task:
  - Nodes generating alpha flows have static public IP addresses, and create repeated alpha flows
  - Can leverage this fact in an offline HNTES design (nightly NetFlow analysis for alpha prefix ID determination)
    - Use /24 prefix IDs
    - Use aging parameter of 30 days
    - Collect NetFlow samples at PE routers' inter-domain links (both directions)
- Router configuration task
  - Use WFQ/PQ 2-queue scheduling (alpha and beta) with no policing
  - IDC support requested: rate-unspecified circuits

Summary

- IDC feature requests
  - rate-unspecified circuits
  - set firewall filters w/o new circuit
- Online HNTES: not required and impractical – (port mirroring unscalable)
  - offline HNTES prototype
- Feedback/questions?
  - mvee@virginia.edu or ctracy@es.net